STAT2_HW2

Xingyu Chen

February 26, 2022

Total Score: 22/28

1 Homework #2

See Canvas for HW #2 assignment due date. Total points without optional question: A1(a-b), A2(a-e), B1(a-b), B2(a-e), so 14 parts $*2 = 28$ total points. Optional question, if you choose to grade it: B3(1-5), so 5 parts $*2 = 10$ points. So, if you choose to grade optional question, $28 + 10$ points $= 38$ points.

1.1 A. Theoretical Problems

1.1.1 Problem A.1

Keep in mind that during class, we kept our equations in the $1, \ldots, \mathrm{p} - 1$ space for predictors. We will use this HW to get familiar with the other very common way of doing it.

Matrices and vectors will play an important role for us in linear regression. Let's review some matrix theory as it might relate to linear regression.

Consider the system of linear equations

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \varepsilon_i$$

for $i = 1, \ldots, n$, where n is the number of data points (measurements in the sample), and $j = 1, \ldots, p$, where

1. $p + 1$ is the number of parameters in the model.

2. Y_i is the i^{th} measurement of the response variable.

3. $X_{i,j}$ is the i^{th} measurement of the j^{th} predictor variable.

4. ε_i is the i^{th} error term and is a random variable (often assumed to be $N\left(0, \sigma^2\right)$).

5. β_j are unknown parameters of the model, $(j = 0, \ldots, p)$. We hope to estimate these, which would help us characterize the relationship between the predictors and response.

(a) Write the equation above in matrix vector form. Call the matrix including the predictors X, the vector of $Y_i\mathbf{sY}$, the vector of parameters β, and the vector of error terms ε. (This is more LaTeX practice than anything else. . .) Q1: 2/2

Answer:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix} = \begin{bmatrix} x_{1,0} & x_{1,1} & \cdots & x_{1,j} \\ x_{2,0} & x_{2,1} & \ddots & x_{2,j} \\ & \vdots & & \vdots \\ x_{i,0} & x_{i,1} & \cdots & x_{i,j} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \end{bmatrix}$$

(b) In class, we will find that the OLS estimator for β in MLR is $\widehat{\beta} = \left(X^T X\right)^{-1} X^T \mathbf{Y}$.

1. What condition must be true about the columns of X for the "Gram" matrix $X^T X$ to be invertible?

   Answer:

   The columns of X are linearly independent.

2. What does this condition mean in practical terms?

   Answer:

   Individuals (observations) are independent. Each features are independent each other. No column can be written as a linear combination of the others. X has full rank.

3. Suppose that the number of measurements (n) is less than the number of model parameters $(p + 1)$. What does this say about the invertibility of $X^T X$? What does this mean on a practical level?

   Answer:

   some variables are removed from the model, either because they are constant or because they belong to a block of collinear variables. The theoretical limit is n-1, as with greater values the X'X matrix becomes non-invertible. Or in other words, the columns of X are not linearly independent.

4. What is true about about $\widehat{\beta}$ if $X^T X$ is not invertible? Q2: 1/2

   Answer:

   We can not find N different values of $\widehat{\beta}$. We know longer have a system of linearly independent equations because X is no longer full rank.

1.1.2 Problem A.2

In class, we defined the hat or projection matrix as

$$H = X \left(X^T X\right)^{-1} X^T$$

The goal of this question is to use the hat matrix to prove that the fitted values, $\widehat{\mathbf{Y}}$, and the residuals, $\widehat{\varepsilon}$, are uncorrelated. We will do it in steps, and some of the proofs will only be required for STAT 5010 students. Note that STAT 4010 students are asked to answer part (e), as to why this result has practical importance.

(a) Show that $\widehat{Y} = HY$. That is, H "puts a hat on" Y. Q3: 1/2

Answer:

Remember that when the coefficient vector is β, the point predictions (fitted values) for each data point are $\mathbf{X}\beta$. Thus the vector of fitted values is

$$\widehat{\mathbf{Y}} \equiv \widehat{\mathbf{m}(\mathbf{X})} \equiv \widehat{\mathbf{m}} = \mathbf{X}\widehat{\beta}$$

Using our equation for $\widehat{\beta}$, we then have

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

where

$$\mathbf{H} \equiv \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$

(b) Show that H is symmetric: $H = H^T$. Q4: 2/2

Answer:

The $(X'X)^{-1}$ is symmetric, then by definition

$$\text{transpose }\left[(X'X)^{-1}\right] = (X'X)^{-1}$$

and thus:

$$\text{transpose }\left[X\,(X'X)^{-1}\,X'\right] = X\,(X'X)^{-1}\,X$$

(c) Show that $H\,(I_n - H) = 0_n$, where 0_n is the zero matrix of size $n \times n$. Q5: 2/2

Answer:

The vector of residuals, $\mathbf{e}$, is

$$\mathbf{e} \equiv \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

the expected residual vector is zero:

$$\mathbb{E}[\mathbf{e}] = (\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \mathbb{E}[\epsilon]) = \mathbf{X}\beta - \mathbf{X}\beta = 0$$

(d) Stating that $\widehat{\mathbf{Y}}$ is uncorrelated with $\widehat{\varepsilon}$ is equivalent to showing that these vectors are orthogonal.* That is, we need to show that their dot product is zero:

$$\widehat{\mathbf{Y}}^T\widehat{\varepsilon} = 0$$

Prove this result. Q6: 1/2

Answer:

$$\sum \widehat{\mathbf{Y}}^T\widehat{\varepsilon} = \sum \widehat{\mathbf{Y}}\widehat{\varepsilon} = \sum(\widehat{\beta_0} + \widehat{\beta_1}X_i)\widehat{\varepsilon} = \sum \widehat{\beta_0}\widehat{\varepsilon} + \widehat{\beta_1}X_i\widehat{\varepsilon} = 0 + 0 = 0$$

(e) Why is this result important in the practical use of linear regression? Q7: 1/2

Answer:

Because $\widehat{\mathbf{Y}}^T\widehat{\varepsilon} = 0$, the residuals satisfy $\text{rank}(X) = p + 1$ linear equalities. Hence, although there are n of them, they are effectively $n - p - 1$ of them. The number $n - p - 1$ is therefore referred to as the degrees of freedom of the residuals $\hat{e}_1, \ldots, \hat{e}_n$.

1.2 B. Computational Problems

1.2.1 Problem B.1

Let $X_1, \dots, X_{30} \overset{\text{iid}}{\sim} N(1, 9)$. The formula for a 90% confidence interval for μ is

$$\bar{X} \pm 1.64 \frac{\sigma}{\sqrt{n}}$$

Let's conduct a simulation to confirm the coverage of this confidence interval.

(a) Generate $m = 500$ random samples of size $n = 30$ from $N(1, 9)$ and calculate the **90%** confidence interval for each. Don't print anything. Q8: 2/2

Answer:

```
confidence_interval <- function(vector) {
  # Standard deviation of sample
  interval = 0.9
  vec_sd <- sd(vector)
  # Sample size
  n <- length(vector)
  # Mean of sample
  vec_mean <- mean(vector)
  # Error according to t distribution
  error <- qt((interval + 1)/2, df = n - 1) * vec_sd / sqrt(n)
  # Confidence interval as a vector
  result <- c("lower" = vec_mean - error, "upper" = vec_mean + error)
  return(result)
}

set.seed(42)                      #set a seed
result = matrix(data = rnorm(30, mean=1, sd=9), nrow = 500, ncol =30)
df <- data.frame(result)
df_result <- data.frame(t(apply(df, 1, confidence_interval)))
```

(b) Estimate the coverage by finding the number of intervals that cover the true mean, and dividing my m. Q9: 2/2

Answer:

```
coverage <- function(x, a, b) {
  a = x['lower']
  b = x['upper']
  if (1 < a | b < 1) {
    return(0)
  }
  else{
    return(1)
  }
}
mean(apply(df_result, 1, coverage))
```

```
## [1] 0.4
```

1.2.2 Problem B.2

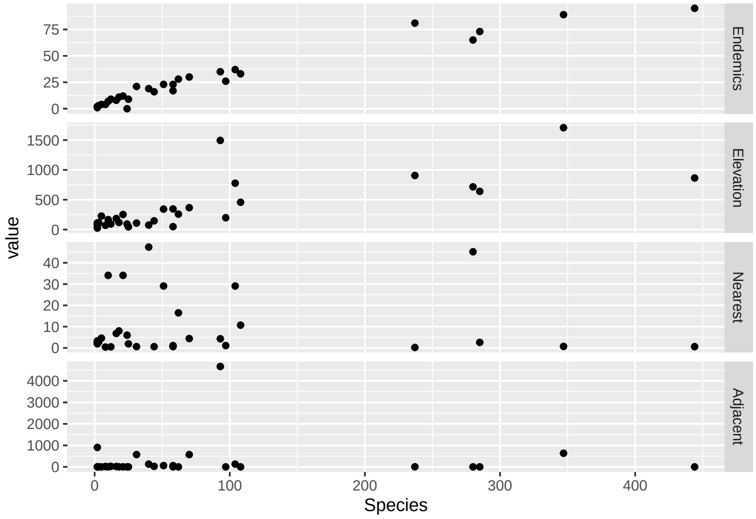
(a) Load the "gala" dataset, and describe the variables. Q10: 2/2
Answer:

```
#install.packages('faraway')
library(faraway)
data("gala")
summary(gala)
```

```
##     Species          Endemics          Area            Elevation
##  Min.   :  2.00   Min.   : 0.00   Min.   :   0.010   Min.   :  25.00
##  1st Qu.: 13.00   1st Qu.: 7.25   1st Qu.:   0.258   1st Qu.:  97.75
##  Median : 42.00   Median :18.00   Median :   2.590   Median : 192.00
##  Mean   : 85.23   Mean   :26.10   Mean   : 261.709   Mean   : 368.03
##  3rd Qu.: 96.00   3rd Qu.:32.25   3rd Qu.:  59.237   3rd Qu.: 435.25
##  Max.   :444.00   Max.   :95.00   Max.   :4669.320   Max.   :1707.00
##     Nearest          Scruz           Adjacent
##  Min.   : 0.20   Min.   :  0.00   Min.   :   0.03
##  1st Qu.: 0.80   1st Qu.: 11.03   1st Qu.:   0.52
##  Median : 3.05   Median : 46.65   Median :   2.59
##  Mean   :10.06   Mean   : 56.98   Mean   : 261.10
##  3rd Qu.:10.03   3rd Qu.: 81.08   3rd Qu.:  59.24
##  Max.   :47.40   Max.   :290.20   Max.   :4669.32
```

(b) Use ggplot() to explore the relationship between the Species variable (response) and Endemics, Elevation, Nearest, and Adjacent (predictor variables). You might do so by creating four separate scatter plots. Do these relationships look linear? Does the variability in Species change as a function of any of the predictors? Are there any outliers in any of the plots? Q11: 2/2
Answer:

```
library(tidyverse)
library(reshape2)
dat <- gala[,c('Species', 'Endemics', 'Elevation', 'Nearest', 'Adjacent')]
dat.m <- melt(dat, id.vars = 'Species')
ggplot(dat.m, aes(Species, value)) +
  geom_point() +
  facet_grid(variable ~ ., scales = 'free')
```



The Endemics and Elevation looks linear and the variability in Species change according to these two variables. There is outliers in adjacent.

(c) Perform a linear regression with Species as the response and Endemics, Elevation, Nearest, and Adjacent as predictors. Interpret the parameter estimate associated with Endemics (assume, for the moment, that the model is correct, and so the interpretation holds). Q12: 2/2

Answer:

```
lmod <- lm(Species ~ Endemics + Elevation + Nearest + Adjacent, gala)
summary(lmod)
```

```
##
## Call:
## lm(formula = Species ~ Endemics + Elevation + Nearest + Adjacent,
##     data = gala)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.232 -10.318   3.412   9.521  70.768
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.510039   8.333335  -2.101   0.0459 *
## Endemics      4.192551   0.429056   9.772  5.1e-10 ***
## Elevation    -0.008594   0.032903  -0.261   0.7961
## Nearest      -0.203743   0.376472  -0.541   0.5932
## Adjacent     -0.005629   0.009876  -0.570   0.5738
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.59 on 25 degrees of freedom
## Multiple R-squared:  0.9464, Adjusted R-squared:  0.9378
## F-statistic: 110.3 on 4 and 25 DF,  p-value: 1.673e-15
```

for every 1 unit increase in Species, the endemics goes up by 4.192551. Can vary by 0.429056. Pr(>|t|) represents the p-value associated with the value in the t value column. If the p-value is less than a certain significance level (e.g. 0.05) the the predictor variable is said to have a statistically significant relationship with the response variable in the model. Here is 5.1 * 10 ^-10 so Endemics have a statistically significant relationship with the Species in the model.

(d) Calculate the residual sum of squares, and the total sum of squares for this model. Then, use these calculations to verify the Multiple R-squared calculation in the summary from the previous part. Interpret R^2 (assume, for the moment, that the model is correct, and so the interpretation holds). Q13: 2/2

Answer:

```
nullmod <- lm(Species ~ 1, gala)
anova(nullmod, lmod)
```

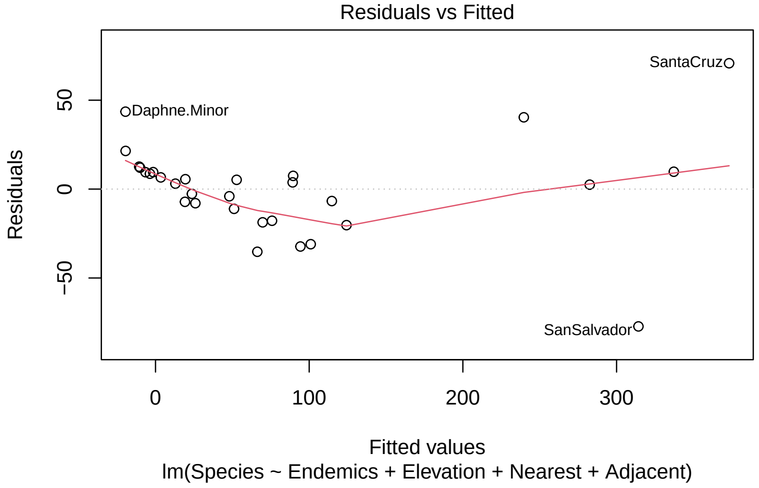
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
29	381081.4	NA	NA	NA	NA
25	20428.8	4	360652.6	110.3383	0

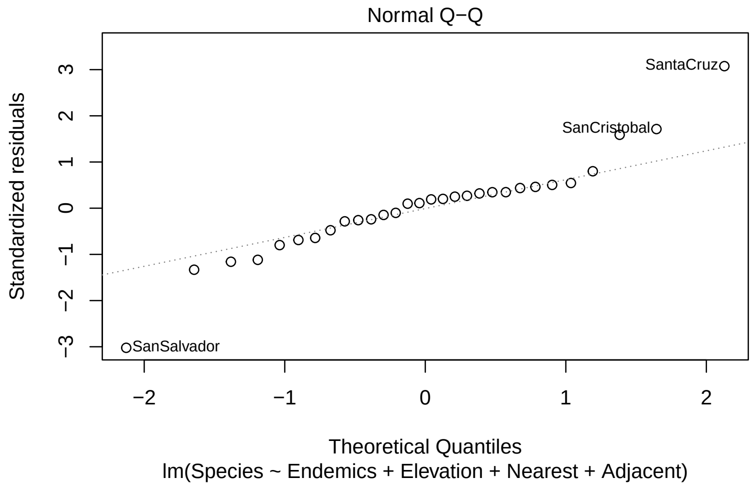
R^2 represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Approximately 94% of the observed variation can be explained by the model's inputs.

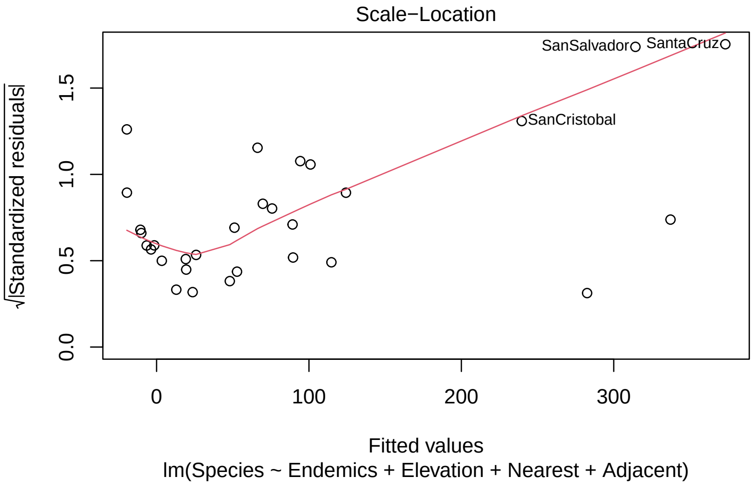
(e) Plot the residuals vs the fitted values. Based on what we've discussed in class (and a question from Section A of this homework!), what do you expect to see in this plot? Do you see what you expect to see? If not, what does that mean? Q14: 1/2

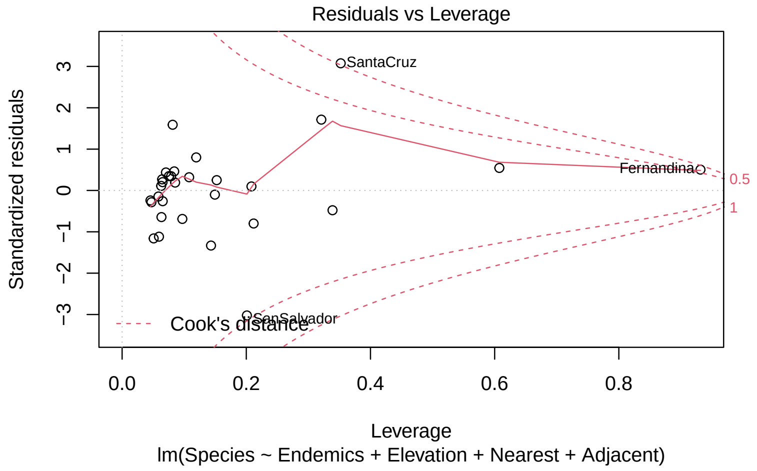
Answer:

```
plot(lmod)
```









It is a scatter plot of residuals on the y axis and fitted values (estimated responses) on the x axis. This plot is used to detect non-linearity, unequal error variances, and outliers.